Literaturgeschichtsschreibung datenbasiert und wikifiziert?
Automatische Extraktion thematischer Statements aus französischen Primärtexten mithilfe von Topic Modeling, RDF und eines kontrollierten Vokabulars in LOD

Röttgermann, Julia
roettger@uni-trier.de
Universität Trier, Germany

Klee, Anne
klee@uni-trier.de
Universität Trier, Germany

Hinzmann, Maria
hinzmannm@uni-trier.de
Universität Trier, Germany

Schöch, Christof
schoech@uni-trier.de
Universität Trier, Germany

Welche Formalisierungs- und Modellierungsarbeit ist nötig, um Kulturen des kollektiven Gedächtnisses wie die Literaturgeschichtsschreibung als Daten abfragbar zur Verfügung zu stellen? Wir sehen aktuell einige Umbrüche in den Strategien der Gedächtnisinstitutionen, die sich zunehmend dem 'Linked Open Data'-Paradigma verpflichtet sehen.[1] Am Beispiel der Domäne französischer Literatur des 18. Jahrhundert verfolgt das Projekt "Mining and Modeling Text" einen ähnlich gearteten, jedoch im Bereich der Literaturgeschichtsschreibung neuen Ansatz einer datenbasierten, wikifizierten Arbeitsweise. Durch den Fokus auf eine spezifische literaturgeschichtliche Domäne entsteht ein besonders dichtes Netz von Aussagen, die über eine systematische Ontologie literaturhistorisch relevanter Aussagetypen miteinander verknüpft sind.

Anhand von drei verschiedenen Informationsquellen (Primärliteratur, Sekundärliteratur und bibliographische Daten) werden literaturhistoriographische Aussagen extrahiert und in Form eines Wissensnetzwerks modelliert, das die heterogenen Daten integriert und über einen SPARQL-Endpunkt abfragbar macht. Das interdisziplinäre Projekt vereint informationswissenschaftliche, juristische, literaturwissenschaftliche und computerlinguistische Expertise. Den Prinzipien der Open Science verbunden, wurde eine Infrastruktur aufgebaut, die freie Software wie *Wikibase* nutzt und im Sinne des FAIRen Datenmanagements die laufenden Ergebnisse transparent auf GitHub zur Verfügung stellt. Doch wie gelingt es, das gewonnene Wissen aus heterogenen Datenquellen so zu modellieren, dass es einander ergänzend und vergleichbar in ein gemeinsames Wissensnetz einfließen kann?

Dies soll zunächst konkret anhand der Extraktion thematischer Aussagen mittels Topic Modeling (1.) sowie der Extraktion von Themenaussagen aus bibliographischen Daten (2.) veranschaulicht werden. Die Relevanz eines kontrollierten Vokabulars im Sinne der Vergleichbarkeit thematischer Aussagewerte unterschiedlicher Informationsquellen (3.) in einem über SPARQL abfragbaren LOD-Wissensnetzwerk (4.) wird im Anschluss dargelegt. Das kontrollierte Vokabular resultiert dabei aus einem Grundstock zeitgeschichtlich relevanter Themen, erweitert um Themenkonzepte, die sich aus der Informationsextraktion der drei Datenquellen ergeben.

Informationsextraktion aus Primärtexten mithilfe von Topic Modeling

Als Datengrundlage zur Modellierung von literaturhistorisch relevanten Aussagen dienen uns drei Kategorien an Texten: Primärliteratur (Romane), Sekundärliteratur (Fachliteratur) und bibliographische Quellen. Die erste der drei Informationsquellen besteht aus einem Korpus aus französischen Romanen der zweiten Hälfte des 18. Jahrhunderts (Röttgermann 2021). Dieses umfasst derzeit 115 Texte und wird laufend durch Volltextdigitalisierung mit dem auf historische Drucke spezialisierten OCR-Tool *OCR4all* (vgl. Reul et al. 2019) und durch Transformation frei verfügbarer EPUBS zu XML-TEI erweitert.[2]

Alle Input-Dateien wurden in TEI-konformes XML nach den Richtlinien der Text Encoding Initiative (vgl. Burnard 2014) nach dem Schema der *European Literary Text Collection* (ELTeC) kodiert (Burnard/Odebrecht 2019). Mithilfe eines Python-Skripts wurden die Texte teilmodernisiert und normalisiert und als Plaintext extrahiert.[3] In dieser Fassung dienen die Texte als Input-Daten für den Topic Modeling-Algorithmus.

Mithilfe von Topic Modeling (vgl. Blei 2011) ist es möglich, "Topics" aus den Primärtexten zu extrahieren, die Aufschluss über Themen-Cluster innerhalb des Korpus geben können. Die Methode generiert auf der Grundlage der Kookkurrenz von Wörtern Gruppen semantisch verwandter Wörtern, die Topics. Hier wurden mit Mallet (vgl. McCallum 2002) zunächst 30 Topics generiert, welchen im Anschluss Label zugewiesen wurden. Diese wiederum wurden mit Konzepten aus einem Themenvokabular verknüpft.[4] Auf diese Weise konnten Themen in den Romanen ermittelt werden, welche auf thematische Konzepte der Epoche der Aufklärung verweisen (vgl. Klee/Röttgermann 2020).

Die gewonnenen Topics verteilen sich folgendermaßen über das Korpus (s. Abb. 1):

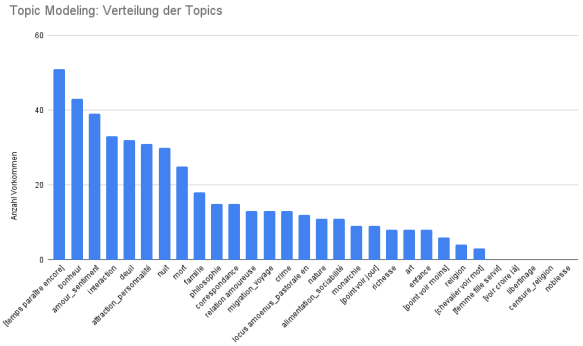

Abb. 1: Dominante Topics im Romankorpus, Topic Modeling mit 92 Romanen und 30 Topics. Daten: DOI: 10.5281/zenodo.4493224.

Neben einigen Topic Modeling Artefakten wie [temps paraître encore] haben sich Topics im Romankorpus herauskristallisiert, die auf literarische Gattungen des 18. Jahrhunderts wie Briefroman (Topic “correspondance”) oder Reiseroman (Topic “migration_voyage”) oder auf häufig thematisch verhandelte Konzepte wie Philosophie (Topic “philosophie”) oder Erziehung (Topic “éducation_enfance”) hinweisen.

Für das Werk Candide (1759) von Voltaire konnte beispielsweise ein Topic extrahiert werden, das mit dem Label “migration_voyage” versehen wurde (Abb. 2).

Abb. 2: Wordle des Topics “migration_voyage” (vgl. Klee/Röttgermann 2020). Daten: https://doi.org/10.5281/zenodo.4493224.

Aus diesem Einzelergebnis leiten wir folgendes angedeutetes Statement über das Werk Candide ab:

[Candide] ABOUT [label @en: travel | label @de: Reise | label@fr: voyage].

Ein Mapping der Entitäten auf Wikidata ergibt daraufhin dieses hier in menschenlesbarer Form angedeutete Statement:

Q215894 [label @fr: Candide] ABOUT Q61509 [label @en: travel | label @de: Reise | label@fr: voyage].

Für den Piloten wurden für das Romankorpus [Stand: 92 Romane][5] je Werk die 5 Topics mit der höchsten Wahrscheinlichkeit als Statements formuliert und so in einem ersten Schritt 460 thematische Aussagen dieser Art generiert. Wir haben zur Einspeisung in das Wissensnetzwerk einen Cutoff von 5 Topics gewählt, um so eine Vergleichbarkeit zur Anzahl der Themen in den bibliographischen Daten herzustellen. Die vollständige Anzahl der Topics pro Werk mit Gewichtungen ist jedoch auf GitHub dokumentiert (Klee/Röttgermann 2020).

Extraktion von thematischen Aussagen aus bibliographischen Daten

Bei der zweiten Informationsquelle handelt es sich um bibliographische Nachweissysteme zur französischen Literatur 1751-1800 (s. Abb. 3). Im Fokus steht *die Bibliographie du genre romanesque français* (Martin et al. 1977), die die Grundgesamtheit der literarischen Produktion der entsprechenden Dekaden sorgfältig dokumentiert und Schlagworte zu thematischen Inhalten der Romane enthält, die jedoch nicht indexiert sind.[6]

Die Bibliographie bietet in Kombination mit den Ergebnissen des Topic Modelings die Möglichkeit eines Mensch-Maschine-Vergleichs – wurden die enthaltenen thematischen Schlagworte doch in den 1970er Jahren durch Lektüre und Zusammentragen von Informationen aus anderen Nachschlagewerken erhoben. Die Bibliographie wurde in mehreren Arbeitsschritten aufwendig erschlossen.[7] Die Extraktion der thematischen Informationen stellt im Wechselspiel mit deren semantischer Modellierung eine besondere Herausforderung dar, da sie einerseits in sehr heterogener Form und andererseits nicht klar abgegrenzt zu weiteren Informationskategorien in der Bibliographie vorliegen.[8] Zur Identifikation der häufigsten thematischen Aussagen, welche als Statements in das Wissensnetzwerk eingespeist werden, wurde das Korpusanalysetool TXM (vgl. Heiden 2010) genutzt.[9] Jede dieser Aussagen wird auf ein Konzept unseres kontrollierten Themenvokabulars gemappt. So können die Strings automatisch extrahiert und als Statements formuliert werden. Aus dem Eintrag zu *Les enfans de la nature* von Pierre Blanchard in der Bibliographie (String aus 4./5. Spalte: <naufrage, robinsonade, intrigue sentimentale; thèmes pédagogiques et philosophiques>) können beispielsweise die folgenden thematischen Statements abgeleitet werden:

[Les enfans de la nature] ABOUT [sentiment | Gefühl | sentiment]

[Les enfans de la nature] ABOUT [pedagogy | Pädagogik | pédagogie]

[Les enfans de la nature] ABOUT [philosophy | Philosophie | philosophie].

Technisch unterscheiden sich die RDF-Triple zu Themen je nach Datenquelle nicht, werden jedoch entsprechend ihrer Herkunft in Wikibase mit der Property *stated in (P14)* referenziert.

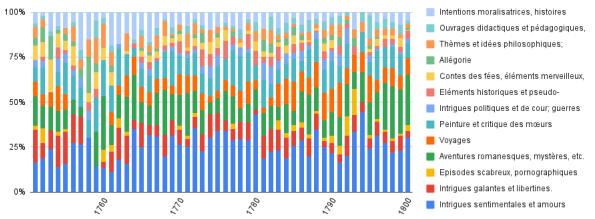

Abb. 3: Themenkategorien des französischen Romans 1751-1800 (Martin et al. 1977: xlviii–xlix).

In den bibliographischen Daten sind insgesamt knapp 2700 Items (Veröffentlichungen fiktionaler Prosa in französischer Sprache inklusive Übersetzungen) enthalten, von denen 349 das thematische Schlagwort ‘voyage’ enthalten.

Rolle des kontrollierten Vokabulars

Wie lassen sich die thematischen Muster in der Primärliteratur mit den Daten aus den bibliographischen Nachweissystemen vergleichen? Ein wichtiger Modellierungsschritt ist zunächst das Erstellen eines kontrollierten Vokabulars aus thematischen Konzepten der französischen Aufklärung, auf das die Ergebnisse des Topic Modeling und die Bibliographie-Schlagworte gemappt werden.

Das Vokabular der Themenbegriffe besitzt eine hohe Relevanz für mehrere Teilprojekte: Es stellt zum einen die Labelbegriffe für die Topics aus dem Topic Model bereit, liefert daneben aber

auch die Konzept-Items für die Objektposition solcher thematischer Statements, die aus der Sekundärliteratur und der Bibliographie extrahiert wurden.

An das Vokabular sind somit mehrere Anforderungen geknüpft: Die Begriffe müssen die Themenkonzepte der französischen Aufklärung abdecken, sollen ein gewisses Abstraktionslevel aufweisen, damit sie als kategorische Begriffe fungieren können und die Zusammenstellung der Begriffe sollte transparent und nachvollziehbar sein. Eine erste Grundlage bildet das Themeninventar des *Dictionnaire européen des Lumières* (Delon et al. 2007). Die Artikelstichwörter bieten eine gute Abdeckung an gesellschaftlich, politisch, ideengeschichtlich oder kulturell relevanten Themen der Epoche und stellen somit einen geeigneten Grundstock an möglichen Labeln für die in den Romanen vorkommenden Themen. Dennoch enthält die Ressource Begriffe, die entweder zu spezifisch (z.B. ‘pyrrhonisme’) oder zu generisch (z.B. ‘fonction’’) sind, um durch sie literarische Themen zu beschreiben, weshalb diese für das Vokabular nicht berücksichtigt wurden. Ergänzt wurden die Begriffe um solche Themenkonzepte, die bei der manuellen Annotation der Sekundärliteratur zusätzlich aufgedeckt wurden, um fehlende Konzepte beim Labeling der Topics sowie um thematische Schlagworte aus der Bibliographie (vgl. Martin et al. 1977), wenn diese anderweitig nicht repräsentiert waren. Das Vokabular ist nun konsolidiert, kann aber auch in Zukunft bei Bedarf erweitert werden.

Um die multilinguale Vergleichbarkeit zwischen französischsprachigen Primärtexten und deutschsprachiger Sekundärliteratur zu gewährleisten, und im Sinne der Anschlussfähigkeit an und Interoperabilität mit anderen Datenbeständen, werden die Themenkonzepte auf einen Normdatensatz (Wikidata) gemappt, wodurch das kontrollierte Vokabular konsolidiert und multilingual erfasst ist (siehe Abb. 4).[10]

https://www.wikidata.org/wiki/Q13648784	terreur	terror	Terror	DEL
https://www.wikidata.org/wiki/Q11635	théâtre	theater	Theater	DEL
https://www.wikidata.org/wiki/Q276183	Théodicée	theodicy	Theodizee	DEL
https://www.wikidata.org/wiki/Q34178	théologie	theology	Theologie	DEL
https://www.wikidata.org/wiki/Q183225	tolérance	Toleration	Toleranz	DEL
https://www.wikidata.org/wiki/Q82821	Tradition	tradition	Tradition	DEL
https://www.wikidata.org/wiki/Q7553	Traduction	translation	Übersetzung	DEL
https://www.wikidata.org/wiki/Q80930	tragédie	tragedy	Tragödie	DEL
https://www.wikidata.org/wiki/Q7590	transport	transport	Verkehr	DEL
https://www.wikidata.org/wiki/Q6958747	Travail	work	Arbeit	DEL
https://www.wikidata.org/wiki/Q186370	troubadour	troubadour	Troubadour	BGRF
https://www.wikidata.org/wiki/Q22082330	tyrannie	tyranny	Tyrannei	Seklit
https://www.wikidata.org/wiki/Q875797	universaux	universality	Universalismus	Seklit
https://www.wikidata.org/wiki/Q59950	Urbanisme	urbanism	Urbanistik	DEL
https://www.wikidata.org/wiki/Q160590	utilitarisme	utilitarianism	Utilitarismus	DEL
https://www.wikidata.org/wiki/Q131156	utopie	utopia	Utopie	DEL
https://www.wikidata.org/wiki/Q194112	valeur	value	Wertvorstellung	Seklit
https://www.wikidata.org/wiki/Q1321250	vanité	vanity	Eitelkeit	Seklit
https://www.wikidata.org/wiki/Q7949	vérité	truth	Wahrheit	Seklit
https://www.wikidata.org/wiki/Q157811	vertu	virtue	Tugend	DEL
https://www.wikidata.org/wiki/Q1129653	vie quotidienne	everyday life	Alltag	Seklit
https://www.wikidata.org/wiki/Q532	village	village	Dorf	DEL
https://www.wikidata.org/wiki/Q515	ville	city	Stadt	DEL
https://www.wikidata.org/wiki/Q124490	violence	violence	Gewalt	Seklit
https://www.wikidata.org/wiki/Q505748	volcan	volcano	Vulkanismus	DEL
https://www.wikidata.org/wiki/Q364340	volonté	will	Wille	Seklit
https://www.wikidata.org/wiki/Q61509	voyage	travel	Reise	DEL
https://www.wikidata.org/wiki/Q431	zoologie	zoology	Zoologie	DEL

Abb. 4: Ausschnitt aus dem kontrollierten Vokabular zur Extraktion thematischer Konzepte. “DEL” bezeichnet dabei Ressourcen aus dem *Dictionnaire européen des Lumières* (Delon et al. 2007), “BGRF” Ressourcen aus den thematischen Schlagworten der *Bibliographie du genre romanesque français, 1751-1800* (Martin et al. 1977).

Unser Ziel: ein Wissensnetzwerk der Literaturgeschichtsschreibung

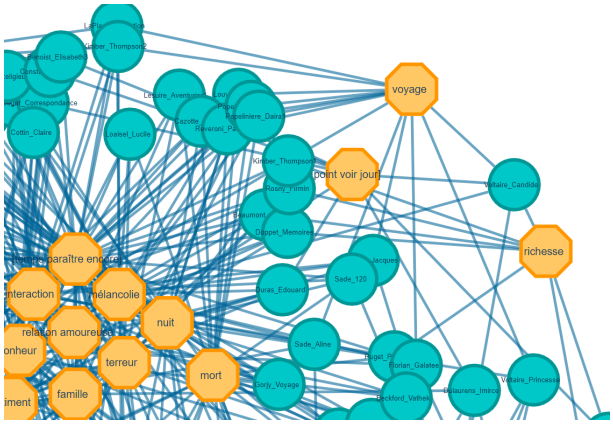

Abb. 5: Ausschnitt aus dem Netzwerk der dominanten Topics pro Werk in *Cytoscape* (Datensatz: Klee/Röttgermann 2020). Siehe https://github.com/MiMo-Text/topicmodeling/blob/master/cytoscape/topics-und-werke_v1.svg für das vollständige Netzwerk.

Ziel ist es, eine Vielzahl an Statements zu aggregieren und die Items auf vielfältige Weise miteinander in Beziehung zu setzen, sodass durch zunehmende Skalierung der Daten aus einzelnen Subjekt-Prädikat-Objekt Aussagen in RDF (Resource Description Framework) ein dichter “Knowledge Graph” (vgl. zum Begriff: Ehrlinger/Wöß 2016) entsteht (s. Abb. 5).

Dieser Graph lässt sich sodann auch über einen SPARQL-Endpoint abfragen (s. Abb.6). Ausgehend von der Beobachtung, dass das Themenkonzept "Reise" in den Bibliographie-Daten bei immerhin 14,7 % der Einträge vermerkt ist, ließe sich beispielsweise fragen, in welchen Werken auch laut Topic Modeling das mit dem Themenkonzept "Reise" verbundene Topic als dominantes Topic vorkommt.[11] Innerhalb des Korpus der 92 Volltexte ergeben sich 13 Treffer.

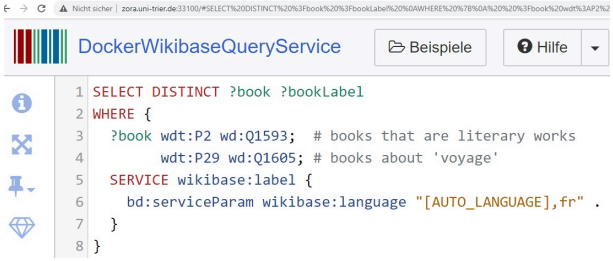

Abb. 6: Projektinterner SPARQL-Endpoint.

Durch Topic Modeling wurden dabei sowohl Werke identifiziert, die auch laut Bibliographie-Daten die Themenkategorie “Voyages” enthalten, als auch solche, in denen dieses Schlagwort nicht vorkam.

Für das Werk *Jacques le fataliste* (1778) von Diderot stimmen Bibliographie-Daten und Topic Modeling-Ergebnisse im Hinblick auf das Themenkonzept “voyage” überein.

98.85 **RÉVÉRONI SAINT-CYR, baron Jacques-Antoine de**
Pauliska ou la perversité moderne, mémoires récens d'une Polonaise
Paris, Lemierre, an VI, 2t. in–12, iv + 254, 222p.
BN†
JP 29 prair. VI; JT 16 prair., 30 fruct. VI, 1 brum. VII; Veillées I 10 1798 55-67, 11 1798 88-91
Gay Monglond Saint-Fargeau
1re personne (mémoires), avec récit intercalé 1re personne; Pologne, Piémont, Suisse; le chevalier de Morsall, madame de Visbourg, Ernest Pradislas, M. d'Olnitz; les malheurs d'une réfugiée, éléments d'actualité.
Autre édition:
– Paris, Lemierre, Pigoreau, Raphel et Bertrandet, an VII, 2t. in–18, 222, 193p. (Tricotel)

Abb. 7: Bibliographie-Eintrag zu *Pauliska ou la perversité moderne* (1798) von Révéroni Saint-Cyr (inhaltliche Schlagworte kursiv).

Für das Werk *Pauliska ou la perversité moderne* (1798) von Révéroni Saint-Cyr wurde mittels Topic Modeling jenes Topic als relevant identifiziert, das mit den Themenkonzept-Labeln "voyage" und "migration" verknüpft ist (Topic Label "migration_voyage"). "Voyage" wird jedoch in den Bibliographie-Daten nicht explizit als Thema genannt (s. Abb. 7). Die Handlungszusammenfassung "les malheurs d'une réfugiée" ("die unglücklichen Begebenheiten einer Flüchtenden") in den bibliographischen Schlagworten (extrahierbar durch das Mapping mit 'fuite' und 'migration' aus dem kontrollierten Themenvokabular) macht jedoch deutlich, dass das Topic "migration_voyage" hier ein plausibles Ergebnis darstellt, das durch den Abgleich mit den bibliographischen Daten präziser eingeordnet und im Wissensnetzwerk verknüpft werden kann. Die semantische Schnittmengen 'menschliche Bewegung' in allen drei Labels ('voyage', 'fuite', 'migration') zeigt die Gratwanderung, die im Herstellen vergleichbarer Themenkonzepte liegt. Das Labeling des Topics als 'migration_voyage' ermöglicht die Vergleichbarkeit der Werke, in denen 'voyage' ein Thema ist und stellt zugleich als Label lediglich eine Annäherung an das oben (vgl. Abb. 2) präziser repräsentierte Topic dar.

SPARQL-Abfragen zu den Ergebnissen des Topic Modelings und/oder der bibliographischen Schlagworten ermöglichen es, auch weniger bekannte Werke zu spezifischen Themen zu ermitteln. Zudem zeichnen sich Muster an Themenkomplexen im Zeitverlauf ab. Für das Thema "voyage" innerhalb der bibliographischen Daten zeigt sich eine (auch statistisch signifikante) ansteigende Entwicklung (vgl. Abb. 8).

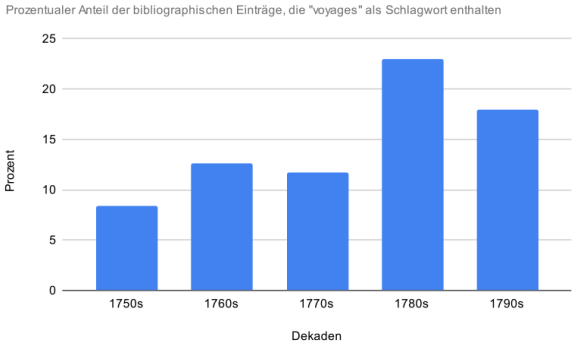

Abb. 8: Prozentualer Anteil der Werke aus der *Bibliographie du genre romanesque français 1751-1800* (Martin et al. 1977), die das Schlagwort "voyages" enthalten.

Eine Erklärung für den Anstieg der Themenkategorie "voyage" in den 1780er und 1790er Jahren könnte sein, dass viele Autor:innen die Handlung ihrer Werke aus politischen Gründen in andere Länder "verlegen" und zudem, dass der Themenkomplex der Reise im Kontext von Emigration im Zuge der politischen Ereignisse zunehmend in Romanen verhandelt wird.

Beispiele hierfür wären *Le roi Guiot* (1791) von Jean Vesque de Puttelange, dessen Protagonist in ferne Königreiche reist, um dort einen neugierigen Blick auf von der absolutistischen Monarchie abweichende politische Herrschaftssysteme zu werfen oder auch der erwähnte Roman *Pauliska ou la perversité moderne* (1798), in dem die Protagonistin durch Polen irrt. Das Thema Flucht und Emigration verweist auf die politische Realität in Frankreich nach der Französischen Revolution (vgl. Van Crugten-André 2001) und ist Anlass zu Reflexionen über Gesellschaftsformen (vgl. Pageaux 1968: 205–14).

Insgesamt ist das Thema "Reise" derzeit laut Topic Modeling im Romankorpus in 14,13% der Werke als dominantes Topic vertreten, in den bibliographischen Daten in 14,74% der Werke. Ein makrostruktureller Blick zeigt demnach in diesem Beispiel eine vergleichbare Größenordnung der thematischen Aussage über das gesamte Korpus hinweg, auch wenn in der Bewertung der Einzelwerke nicht immer Kongruenz besteht.

Fazit und Perspektiven

Das Projekt MiMoText modelliert die Geschichte des französischen Romans der zweiten Hälfte des 18. Jahrhunderts in Form von RDF-Tripeln in Wikibase als Knowledge Graph. Im Zuge eines Pilotprojekts wurden in einem ersten Schritt aus bibliographischen Daten und aus einem Romankorpus Relationen zwischen Werken und Themen extrahiert, die in Form von Tripeln in eine eigene Wikibase-Instanz eingelesen wurden. Zur Modellierung der Themen des französischen Romans der Aufklärung wurde ein kontrolliertes Vokabular erstellt, welches auf Wikidata gemappt wurde, um anschlussfähig an die Linked Open Data Cloud zu sein. Die Ergebnisse der Informationsextraktion aus den Romanen (mithilfe von Topic Modeling) und der Informationsextraktion aus bibliographischen Daten können nun per SPARQL-Endpoint abgefragt werden.

Zu den nächsten Schritten gehört neben der Extraktion weiterer Statements über quantitative Romananalysen der Import von Themen-Statements aus dem dritten Typus von Informationsquellen (Fachliteratur) in unsere Wikibase-Instanz.[12]

Fußnoten

1. Als Beispiel sei hier die Initiative der GND genannt, die eigenen Daten in Wikidata oder zumindest in einer Wikibase-Instanz (die Software hinter Wikidata) zu integrieren: https://blog.wikimedia.de/2020/03/04/wikibase-und-gnd/, letzter Zugriff: 30.11.2021. Zum Begriff "Linked Open Data" vgl. (Berners-Lee et al. 2006: 1–130).
2. Wir nutzen die Quellen *Wikisource*, *Ebooks libres et gratuits*, *GoogleBooks*, *Rousseau Online* und *Frantext*. Diese Metadatentabelle dokumentiert die Korpuszusammensetzung und wird parallel zum Korpusaufbau laufend aktualisiert: http://doi.org/10.5281/zenodo.5040855 /. https://github.com/MiMoText/roman18/blob/master/XML-TEI/xml-tei_metadata.tsv, letzter Zugriff: 30.11.2021.

3. https://github.com/MiMoText/roman18/blob/master/Python-Scripts/tei2txt_run.py, letzter Zugriff: 30.11.2021.
4. In der Regel werden die Topics durch ein Element des Themenvokabulars repräsentiert, in wenigen Fällen erscheint die Repräsentation durch zwei Themenkonzept-Label treffender.
5. Die Datengrundlage dieses Topic Modeling Durchgangs ist unter folgendem Release zu finden: (Klee/Röttgermann 2020).
6. Zu möglicherweise fehlenden Werken vgl. Dawson 1978. Dawson benennt auch das Desiderat eines Themenindexes.
7. An das Scannen sowie OCR schlossen sich das Generieren von Trainingsdaten sowie die Auszeichnung aller Einträge über ein Machine Learning-Verfahren (CRF) in XML an. Diese bildeten die Datengrundlage für die anschließende Modellierung der Einträge in RDF (vgl. Lüschow 2020), bei der jedoch die einzelnen Keywords noch nicht semantisch modelliert worden sind.
8. Weitere Kategorien umfassen die Erzählform, den Ort der Handlung, die Figuren des Romans, die Tonalität/den Stil des Werks.
9. Ausgewählt wurden zum einen Strings mit mindestens acht Vorkommen in der Bibliographie und darüber hinaus solche mit einer besonderen Relevanz in den anderen Informationsquellen und für die literaturgeschichtliche Domäne insgesamt wie zum Beispiel der String ‘robinsonade’.
10. Zur Dokumentation der Liste vgl. Klee/Hinzmann 2021.
11. Als dominante Topics bezeichnen wir jeweils diejenigen Topics, die in einem Werk unter den 5 Topics mit den höchsten Wahrscheinlichkeiten sind (siehe oben).
12. Hierfür werden mit INCEpTION Aussagen in literaturgeschichtlichen Fachtexten annotiert, die in das Wissensnetzwerk eingespeist werden und zugleich als Trainingsdaten für die automatische Extraktion von Themenaussagen dienen.

Bibliographie

Berners-Lee, Tim / Hall, Wendy / Hendler, James A. / O’Hara, Kieron / Shadbolt, Nigel / Weitzner, Daniel J. (2006): “A Framework for Web Science”, in: *Foundations and Trends in Web Science* 1.1.: 1–130. 10.1561/1800000001.

Blei, David M. (2011): “Introduction to Probabilistic Topic Models”, in: *Communications of the ACM 55.4.*: 1–16.

Burnard, Lou (2014): *What Is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resource*s. Marseille: Encyclopédie Numérique.

Dawson, Robert L. (1978). “The Martin, Mylne, Frautschi Bibliographie Du Genre Romanesque Français”, in: *Eighteenth-Century Studies*, 11.4., 497–508. 10.2307/2737969.

Delon, Michel (2007): *Dictionnaire européen des Lumières*. Paris: PUF.

Ehrlinger, Lisa / Wöß, Wolfram (2016): „Towards a Definition of Knowledge Graphs“, in: *SEMANTiCS (Posters, Demos, SuCCESS)* http://ceur-ws.org/Vol-1695/paper4.pdf.

Heiden, Serge (2010): *The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme*. http://halshs.archives-ouvertes.fr/halshs-00549764/en [letzter Zugriff: 30.11. 2021].

Klee, Anne / Hinzmann, Maria (2021): *MiMoText/vocabularies* [Data Set]. https://github.com/MiMoText/vocabularies/blob/main/thematic_vocabulary.tsv [letzter Zugriff: 30.11.2021].

Klee, Anne / Röttgermann, Julia (2020): *Doing Topic Modeling on French 18th Century Novels in the Context of MiMoText Project [Data Set]* https://github.com/MiMoText/topicmodeling [letzter Zugriff: 30.11.2021].

Lüschow, Andreas (2020): “Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane”, in: *Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*, 80–84 10.5281/zenodo.3666690.

Martin, Angus / Mylne, Vivienne / Frautschi, Richard L. (1977): *Bibliographie du genre romanesque* français, 1751-1800. London: Mansell.

McCallum, Andrew Kach (2002): *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu.

Pageaux, Daniel-Henri (1968): “Voyages romanesques au siècle des Lumières”, in: *Études littéraires*, 1.2.: 205–214. 10.7202/500020ar.

Reul, Christian / Christ, Dennis / Hartelt, Alexander / Barbelt, Nico / Wehner, Maximilian (2019): “OCR4all -- An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings”, in: *ArXiv:1909.04032* [Cs].

Röttgermann, Julia (2021): *Collection de romans français du dix-huitième siècle (1750-1800) / Eighteenth-Century French Novels (1750-1800) [dataset]. Release v0.2.0* 10.5281/zenodo.5040855.